Types of Tasks	Token Number	Token Type Number	Token Duplication Rate	Vocab Coverage
Mathematical Reasoning				
Arithmetic (primary difficulty)	16136	14	99.9%	0.04%
Arithmetic (middle-school difficulty)	5663	16	99.7%	0.05%
Algebra	5234	107	98.0%	0.33%
Geometry	2615	75	97.1%	0.23%
Counting and probability	2524	43	98.3%	0.13%
number theory	2395	71	97.1%	0.22%
precalculus	3388	84	97.5%	0.26%
Average	*5422*	*58*	*98.9%*	*0.18%*
Text Generation				
Translation	2500	1065	57.4%	3.32%
	5000	1832	63.3%	5.10%
	10000	2980	70.2%	9.31%
Average	*5833*	*1959*	*66.4%*	*6.12%*
Summarization	2500	1265	49.4%	4.01%
	5000	1970	60.6%	6.16%
	10000	3192	68.0%	9.98%
Average	*5833*	*2142*	*63.2%*	*6.69%*

Table 2: Statistics and comparisons of token number, type, duplication rate, vocab coverage on mathematical reasoning (from seven types of different task domains and difficulties), translation, and summarization tasks.

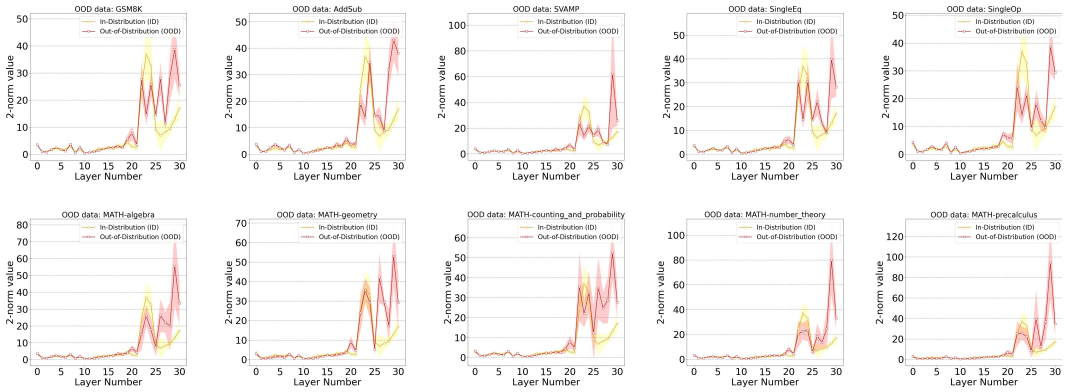

Figure 1: Trajectory volatility (embedding 2-norm differences between neighboring layers) curves and sample standard deviation (color shading) of ID dataset and 10 OOD datasets from different domains and difficulties.

Dataset	1-31 layers	20-31 layers	26-31 layers
ID Dataset			
MultiArith	6.53	14.84	10.89
Near-shift OOD Dataset			
GSM8K	8.60	20.55	26.43
SVAMP	8.02	18.82	24.68
AddSub	8.72	20.54	27.24
SingleEq	8.14	19.26	22.42
SingleOp	7.50	17.17	21.62
Far-shift OOD Dataset			
Algebra	8.83	21.35	31.50
Geometry	10.00	25.27	34.14
CountandProb	10.30	25.77	33.70
NumberTheory	9.40	23.10	33.86
Precalculus	11.06	28.30	43.60

Table 1: The average volatility statistics for layers 1-31 (full layers), 20-31, and 26-31 on each dataset corresponding to Figure 1.